Full Parameter Fine-tuning for Large Language Models with Limited Resources

Year: 2023 | *Citations: 179* | *Authors: Kai Lv, Yuqing Yang, Tengxiao Liu*

Abstract

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) but demand massive GPU resources for training. Lowering the threshold for LLMs training would encourage greater participation from researchers, benefiting both academia and society. While existing approaches have focused on parameter-efficient fine-tuning, which tunes or adds a small number of parameters, few have addressed the challenge of tuning the full parameters of LLMs with limited resources. In this work, we propose a new optimizer, LOw-Memory Optimization (LOMO), which fuses the gradient computation and the parameter update in one step to reduce memory usage. By integrating LOMO with existing memory saving techniques, we reduce memory usage to 10.8% compared to the standard approach (DeepSpeed solution). Consequently, our approach enables the full parameter fine-tuning of a 65B model on a single machine with 8 RTX 3090, each with 24GB memory.Code and data are available at https://github.com/OpenLMLab/LOMO.